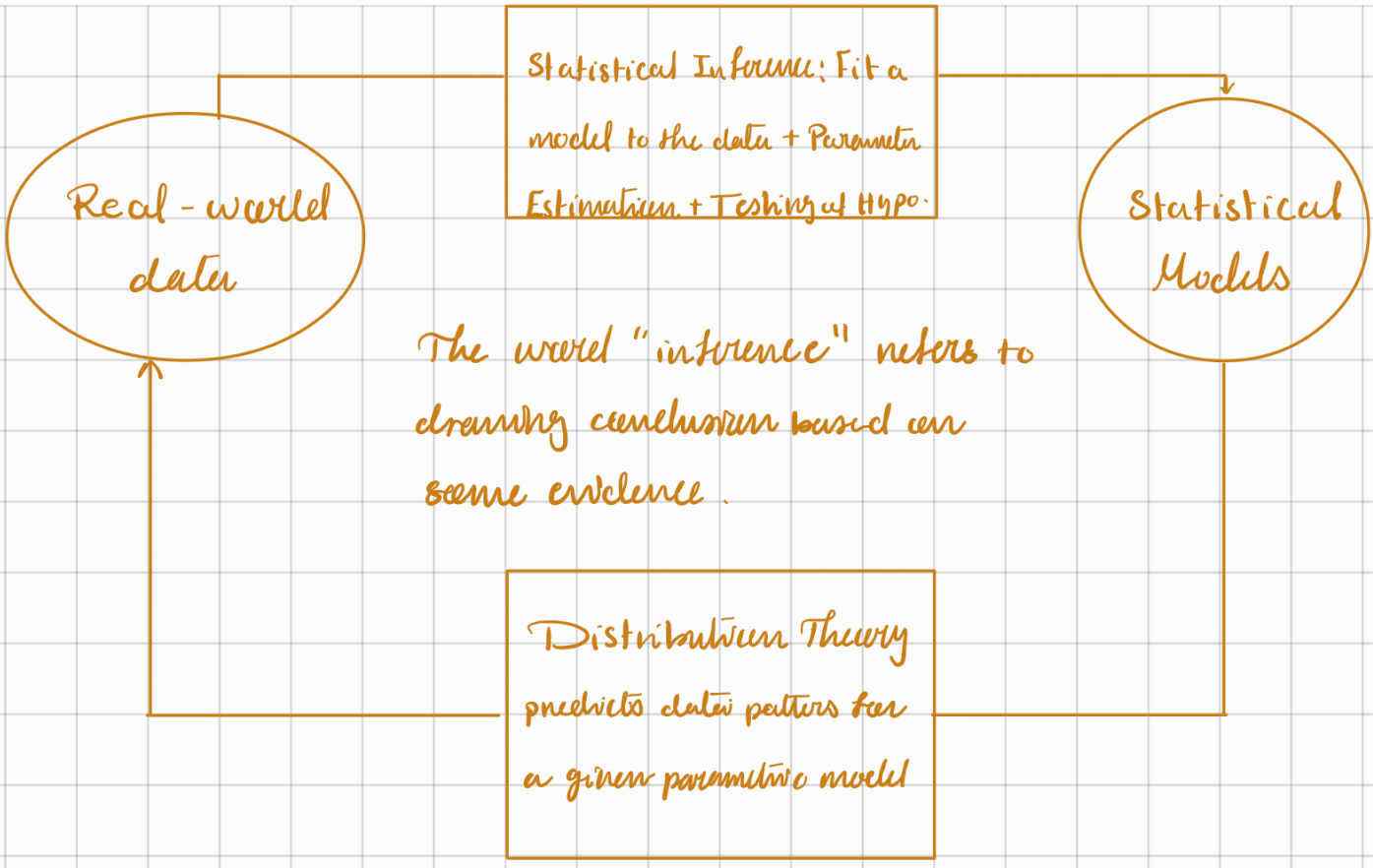

- Statistical Inference refers to drawing conclusion based on evidence obtained from the data.
- Statistical data consists of variability and uncertainty statistics models variability and quantifies uncertainty.

Statistical Inference
- Frequentist
- Bayesian

Parametric
- When the population distribution is known e.g Normal, Poisson....
  - Testing of Hypothesis (verifying θ)
  - Estimation (Estimating θ)

Non-parametric
- When we have no idea about the data generating process. e.g Kernel distribution

(Verifying θ) Point Estimation Interval Estimation

* **Main challenges** :

1) How to summarize the information in the data using formal mathematical tools ? Min Q1 Median Q3 Max , Mean

2) How to use these summaries to answer questions about the phenomenon of interest ?

* **Examples** : 1) Survival analysis 2) Biostatistics
3) NASA dataset produced in their software detect detection trails
4) Daily price of the saleable stocks for a company

* Data : $x_1, \dots, x_n$ where x_j can either be a scalar or a vector.
In stat inference, this sample is interpreted as a realization of RVs.

* Notation : $\tilde{X} = (X_1, \dots, X_n)^T$ is a vector of RVs
$\tilde{x} = (x_1, \dots, x_n)^T$ is the sample of observations

* Statistic : Function of sample that helps in data summarization (reduction)
$T(\tilde{x}) : \mathbb{R}^n \to \mathbb{R}^m$; $1 \leq m \leq n$ e.g $m = 2$ (μ, σ^2) for normal
Instead of reporting the entire sample, only report the value of statistic
Visual 5-point summary statistic : Boxplot

* Point Estimation : Uses sample data to calculate a single value (statistic) which serves as the best estimate of the unknown (fixed / random) population parameter
- $(X_1, \dots, X_n) \sim F_\theta$, $\theta \in \Omega$, Guess θ ? or Guess $\tau(\theta)$?
- $T(\tilde{x})$ is called the point estimate of θ.

* Interval Estimation : Interval of possible (probable) values of an unknown population parameter. $L(\tilde{x}) \leq \theta \leq U(\tilde{x})$

$\theta \in [L(\tilde{x}), U(\tilde{x})]$ is called an interval estimators.

Bayesian : Credible interval (deal with prior info)

Frequentist : Confidence interval / Prediction Intervals

* Parametric Models : $\{f(x|\theta) \mid \theta \in \Omega\}$ with k parameters $\Omega \subseteq \mathbb{R}^k$ is the parameter space to which the parameters belong and $f(x|\theta)$ is the PDF / PMF for the distribution (k is finite)

* Point Estimation : 1) Method of Moments

2) Maximum Likelihood Estimation

* Measure of Goodness : 1) Unbiasedness 3) Efficient

2) Consistency 4) Completeness

* Method of Moments : $X_1, \ldots, X_n \overset{iid}{\sim} f(x|\theta)$. How to estimate θ?

$X \sim f(x|\theta)$; $\theta \in \mathbb{R}^k$ so we consider the first k moments of the distribution of X. To estimate θ, we use observed sample moments

$$\mu_1 = E(X) \qquad \hat{\mu}_1 = \frac{1}{n}(X_1 + \ldots + X_n) = \bar{X}$$

$$\mu_2 = E(X^2) \qquad \hat{\mu}_2 = \frac{1}{n}(X_1^2 + \ldots + X_n^2)$$

$$\vdots \qquad\qquad \vdots$$

$$\mu_k = E(X^k) \qquad \hat{\mu}_k = \frac{1}{n}(X_1^k + \ldots + X_n^k)$$

equate the sample moments to the theoretical moments

$$\hat{\mu}_2 - \hat{\mu}_1^2 = \frac{1}{n}\sum (X_i - \bar{X})^2$$

1) $X \sim Pois(\lambda)$

$E(X) = \lambda \Rightarrow \hat{\lambda} = \frac{1}{n}(X_1 + \dots + X_n) = \bar{X}$

2) $X \sim exp(\lambda)$

$E(X) = \frac{1}{\lambda} \Rightarrow \frac{1}{\hat{\lambda}} = \frac{1}{n}(X_1 + \dots + X_n) \Rightarrow \hat{\lambda} = \frac{1}{\bar{X}}$

3) $X \sim N(\mu, \sigma^2)$

$E(X) = \mu$

$E(X^2) = Var(X) + E(X)^2 = \sigma^2 + \mu^2$

$\hat{\mu}_1 = \mu \Rightarrow \hat{\mu} = \bar{X}$

$\hat{\mu}_2 = \sigma^2 + \mu^2 = \sigma^2 + \hat{\mu}_1^2 \Rightarrow \hat{\sigma^2} = \hat{\mu}_2 - \hat{\mu}_1^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2$

$= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$

4) $X \sim Gamma(\alpha, \beta)$

$E(X) = \frac{\alpha}{\beta} = \hat{\mu}_1$

$E(X^2) = Var(X) + E(X)^2 = \frac{\alpha}{\beta^2} + \frac{\alpha^2}{\beta^2} = \frac{\alpha + \alpha^2}{\beta^2} = \hat{\mu}_2$

$\frac{\hat{\alpha}}{\hat{\beta}} = \bar{X} \Rightarrow \hat{\alpha} = \hat{\beta}\bar{X} = \frac{n\bar{X}^2}{\sum(X_i - \bar{X})^2}$

$\frac{\hat{\alpha} + \hat{\alpha}^2}{\hat{\beta}^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \Rightarrow \frac{\hat{\beta}\bar{X} + \hat{\beta}^2\bar{X}^2}{\hat{\beta}^2} = \frac{1}{n}\sum X_i^2$

$\Rightarrow \bar{X}\left(\frac{1}{\hat{\beta}} + \bar{X}\right) = \frac{1}{n}\sum X_i^2$

$\Rightarrow \frac{1}{\hat{\beta}} + \bar{X} = \frac{1}{n\bar{X}}\sum X_i^2$

$\Rightarrow \frac{1}{\hat{\beta}} = \frac{1}{n\bar{X}}\sum X_i^2 - \bar{X} = \frac{1}{\bar{X}}\left(\frac{1}{n}\sum X_i^2 - \bar{X}^2\right)$

$\Rightarrow \hat{\beta} = \frac{n\bar{X}}{\sum(X_i - \bar{X})^2}$

* Method of MLE : $L(\theta|\vec{x}) = \prod_{i=1}^{n} f(x_i|\theta)$ likelihood function

MLE is finding θ that maximizes $L(\theta|x)$ ($L(\theta|x)$ measures how likely is a particular value of θ given $\vec{x}$ is observed)

$$\hat{\theta} = \underset{\theta \in \Omega}{\arg\max}\,(L(\theta|\vec{x}))$$ Optimization Problem

We can maximize $l(\theta) = \log(L(\theta|\vec{x})) = \sum_{i=1}^{n} \log(f(x_i|\theta))$ instead

1) $X \sim Ber(p)$

$$L(p|x) = \prod_{i=1}^{n} f(x_i|p)$$

$$= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

$$= p^{\sum x_i}(1-p)^{n-\sum x_i}$$

$$l(p) = \sum_{i=1}^{n} x_i \log(p) + (n - \sum_{i=1}^{n} x_i)\log(1-p)$$

$$\frac{\partial l(p)}{\partial p} = \frac{1}{p}\sum_{i=1}^{n} x_i - \frac{1}{1-p}(n - \sum_{i=1}^{n} x_i) \overset{set}{=} 0$$

$$(1-p)\sum x_i - p(n - \sum_{i=1}^{n} x_i) = 0$$

$$\sum_{i=1}^{n} x_i - pn = 0$$

$$\hat{p}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{X}$$

Checking: $\frac{\partial^2}{\partial p} l(p|x) = \frac{\partial}{\partial p}\left(\frac{\sum x_i - np}{p-(1-p)}\right) < 0 \Rightarrow$

$\hat{p} = \bar{X}$ is the MLE of p

2) Let $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$

$$\mathcal{L}(\mu, \sigma^2 \mid X_i) = \prod_{i=1}^{n} f(x_i \mid \mu, \sigma^2)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{(x_i - \mu)^2}{2\sigma^2} \right)$$

$$= (2\pi\sigma^2)^{-n/2} \exp\left(-\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} \right)$$

$$l(\mu, \sigma^2) = \frac{-n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}$$

$$\frac{\partial l(\mu, \sigma^2)}{\partial \mu} = \frac{+1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = \frac{1}{\sigma^2} \left(\sum_{i=1}^{n} x_i - n\mu \right) \overset{set}{=} 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i = n\mu \quad \Rightarrow \hat{\mu}_{MLE} = \bar{x}$$

$$\frac{\partial l(\mu, \sigma^2)}{\partial \sigma^2} = \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \hat{\mu}_{MLE})^2 \overset{set}{=} 0$$

$$\Rightarrow \sigma^2 n = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$\Rightarrow \sigma^2_{MLE} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

3) $X_1, \dots, X_n \sim P(\lambda)$

$$L(\lambda \mid X) = \prod_{i=1}^{n} f(x_i \mid \lambda)$$

$$= \prod_{i=1}^{n} \frac{e^{-\lambda} \lambda^{x_i}}{x_i!}$$

$$= \frac{e^{-n\lambda} \cdot \lambda^{\sum x_i}}{\prod_{i=1}^{n} (x_i)!}$$

$$l(\lambda) = -n\lambda + \log(\lambda) \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \log(x_i!)$$

$$\frac{\partial \ell(\lambda)}{\partial \lambda} = -n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i \overset{\text{set}}{=} 0 \Rightarrow \hat{\lambda}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}$$

4) Exponential : $f(x_i \mid \mu, \lambda) = \lambda\, e^{-\lambda(x_i-\mu)} = \lambda e^{-\lambda x_i + \lambda\mu}$, $x \geq \mu$, $\mu \in \mathbb{R}$, $\lambda > 0$

$$L(x_i \mid \mu, \sigma^2) = \prod_{i=1}^{n} f(x_i \mid \mu, \lambda)$$

$$= \prod_{i=1}^{n} \lambda\, e^{-\lambda(x_i-\mu)}$$

$$= \lambda^n e^{-\lambda \sum_{i=1}^{n}(x_i-\mu)}$$

$$\ell(\mu, \lambda) = n \log \lambda - \lambda \sum_{i=1}^{n}(x_i - \mu)$$

- As e^y is an increasing function then $\forall y, y' ;\ y < y'\ \ e^y < e^{y'}$ so to maximimize such a function we need to consider $\hat{\mu}_{MLE} = \min_{1 \leq i \leq n}(x_i)$ ↓ as $\mu \leq x_i\ \forall i$

- $\frac{\partial \ell(\mu, \lambda)}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n}(x_i - \hat{\mu}) \overset{\text{set}}{=} 0 \Rightarrow$

$$n = \lambda \sum_{i=1}^{n}(x_i - \hat{\mu}) \Rightarrow$$

$$\hat{\lambda}_{MLE} = \frac{n}{\sum_{i=1}^{n} x_i - n x_{(1)}} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} x_i - x_{(1)}} = \frac{1}{\overline{x} - x_{(1)}}$$

5) $X \sim Unif[a, b]$ $\quad f_{x_i} = \frac{1}{b-a}\, \mathbb{1}(a \leq x_i \leq b)$

$$L(a, b \mid X) = \prod_{i=1}^{n} f(x_i \mid a, b)$$

$$= \frac{1}{(b-a)^n}\, \mathbb{1}(a \leq x_{(1)} \leq x_{(n)} \leq b)$$

$$= \frac{1}{(b-a)^n}\, \mathbb{1}(a \leq x_{(1)})\, \mathbb{1}(x_{(n)} \leq b)$$

As $x \mapsto \frac{1}{x}$ is a decreasing function, $\hat{a}_{MLE} = X_{(1)}$ and $\hat{b}_{MLE} = X_{(n)}$

c) $X_1, \dots, X_K \sim \text{Multinomial}(n, (p_1, \dots, p_K))$ (Important for final)

$$f(x_1, \dots, x_K | p_1, \dots, p_K) = \binom{n}{x_1, \dots, x_K} p_1^{x_1} \cdots p_K^{x_K}, \quad \sum_{i=1}^{K} p_i^{x_i} = 1 \text{ and}$$

$$\ell(p_1, \dots, p_K) = \log \binom{n}{x_1, \dots, x_K} + \sum_{i=1}^{K} x_i \log(p_i) \qquad \sum_{i=1}^{K} n_i = n$$

$$F(p_1, \dots, p_K) = \ell(p_1, \dots, p_K) + \lambda g(p_1, \dots, p_K)$$
$$= \log \binom{n}{x_1, \dots, x_K} + \sum_{i=1}^{K} x_i \log(p_i) + \lambda \left(\sum_{i=1}^{K} p_i - 1 \right)$$

$$\frac{\partial F(p_1, \dots, p_K)}{\partial p_i} = \frac{x_i}{p_i} + \lambda = 0 \Rightarrow \hat{p}_i = \frac{-x_i}{\lambda} \qquad \forall i \in \{1, \dots, K\}$$

Choose $\lambda = -n$ then $\forall i \in \{1, \dots, K\}$ $\hat{p}_i = \frac{x_i}{n}$ so $\sum_{i=1}^{K} \frac{x_i}{n} = \frac{n}{n} = 1$ thus this λ verifies the constraint.

Alternatively, $\sum_{i=1}^{K} \hat{p}_i = \sum_{i=1}^{K} \frac{-x_i}{\lambda} = \frac{-n}{\lambda} = 1 \Rightarrow \lambda = -n$

* The Tank Problem:

The enemy has an unknown number N of tanks which he has obligingly numbered 1, 2, ..., N. Spies have reported sighting 8 tanks with numbers

$$\tilde{n} = (137, 24, 86, 33, 92, 129, 17, 111)^T$$

Assume that sightings are independent and that each of the N tanks has a probability $\frac{1}{N}$ of being observed at each sighting. What is the MLE of N?

$$L(N | n) = \begin{cases} \frac{1}{N^8}, & N \geq 137 \\ 0, & N < 137 \end{cases} \qquad \hat{N} = X_{(8)} = 137$$

C_θ , $N \geq 137$

* Why MLE ? 1) The estimators are consistent and asymptotically normal i.e for likelihood $L(\theta | n)$ & n sample size

$\hat{\theta}_{MLE} \longrightarrow \theta$ as $n \to +\infty$

$\sqrt{n}(\hat{\theta}_{MLE} - \theta) \xrightarrow{d} N(0, \Sigma^*)$

Σ^* is an estimate matrix called Fisher information matrix

So if we use MLE method, we can construct confidence interval around $\hat{\theta}_{MLE}$

2) Sometimes the moments do not exist so we cannot use the method of moments.

* Sometimes we don't find a closed form solution of MLE

$$L(\alpha, \beta | X) = \prod_{i=1}^{n} f(x_i | \alpha, \beta)$$

$$= \prod_{i=1}^{n} \frac{\beta^\alpha}{\Gamma(\alpha)} x_i^{\alpha-1} \cdot e^{-\beta x_i}$$

$$= \frac{\beta^{n\alpha}}{\Gamma(\alpha)^n} \left(\prod_{i=1}^{n} x_i^{\alpha-1} \right) e^{-\beta \sum_{i=1}^{n} x_i}$$

$$l(\alpha, \beta) = n\alpha \log(\beta) - n \log(\Gamma(\alpha)) + (\alpha - 1) \sum_{i=1}^{n} \log(x_i) - \beta \sum_{i=1}^{n} x_i$$

$$\frac{\partial l(\alpha, \beta)}{\partial \beta} = \frac{n\alpha}{\beta} - \sum x_i = 0 \Rightarrow n\alpha = \beta \sum_{i=1}^{n} x_i$$

$$\Rightarrow \hat{\beta}_{MLE} = \frac{\hat{\alpha}}{\bar{x}}$$

$$\frac{\partial l(\alpha, \beta)}{\partial \alpha} = n \log(\beta) - n \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum_{i=1}^{n} \log(x_i)$$

$$= n \log\left(\frac{\hat{\alpha}}{\bar{x}}\right) - n \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum_{i=1}^{n} \log(x_i)$$

* We will study optimization methods : 1) Newton Raphson (NR)
2) Gradient Descent (GD)

1) Newton Raphson :

Taylor Expansion : $f(n) = f(\alpha) + (n-\alpha) f'(\alpha) = 0 \Rightarrow$

$-f(\alpha) = (n-\alpha) f'(\alpha) \Rightarrow$

$n = \alpha - \frac{f(\alpha)}{f'(\alpha)}$

We need to start with an initial guess ($\alpha^{(0)}$), which may be the method of moments estimator $\alpha^{(0)} = \frac{n\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$

Then, we compute $\alpha^{(1)}, \alpha^{(2)}, \dots, \alpha^{(n)}$ selection criterion is whether $\alpha^{(i)}$ is better than $\alpha^{(j)}$.

Having computed $\alpha^{(t)}$, $t \in \{0,1,\dots,n\}$, we compute the next iteration α^{t+1} by approximating the function with a linear equation using the 1st order Taylor Expansion around $\hat{\alpha} = \alpha^{(t)}$, and set $\alpha^{(t+1)}$ as the value of $\hat{\alpha}$ that solves this linear equation.

$$0 = \log(\hat{\alpha}) - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} - \log(\bar{x}) + \frac{1}{n}\sum_{i=1}^{n}\log x_i$$

$$f(\alpha) = \log(\alpha) - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} \qquad \theta_{t+1} = \theta_t - \frac{f'(\theta_t)}{f''(\theta_t)}$$

we have the same result as we originally considered $\frac{\partial \ell(\alpha,b)}{\partial \alpha}$ and then applied NR

$\alpha^{(1)} = \alpha^{(0)} - \frac{f(\alpha_0)}{f(\alpha_1)}$ if we think $\alpha^{(1)}$ is better the we update it

$\alpha^{(2)} = \alpha^{(1)} - \frac{f(\alpha_1)}{f'(\alpha_1)}$

2) Gradient Ascent :

Taylor Expansion : $f(\theta) = f(\theta_0) + (\theta-\theta_0) f'(\theta_0) + \frac{(\theta-\theta_0)^2}{2} f''(\theta_0)$

Suppose $f''(\theta_0)$ is -ve constant : $f(\theta) = f(\theta_0) + (\theta - \theta_0) f'(\theta_0) - \frac{(\theta - \theta_0)^2}{2t}$

Hence, we get $\theta = \theta_0 + t f'(\theta_0)$ by setting the derivative to zero

t ↓ Learning Rate

The choice of the learning rate is the most important consideration.

$f'(\theta) = f'(\theta_0) - \frac{(\theta - \theta_0)}{t} = 0$ as $f''(\theta_0) = \frac{-1}{t}$

$\Rightarrow \theta = \theta_0 + t f'(\theta_0)$

t ↓ Tuning Parameter

Concave Function : Reach global maximum

General Function : Reach local maximum

* Higher Dimensions : $\theta_{k+1} = \theta_k + t \nabla f(\theta_k)$

* Bais, Variance and Mean Squared Error :

1) Single Parameter : $X_1, \ldots, X_n \overset{iid}{\sim} f(x|\theta)$. We estimate θ using MOM and MLE ($\hat{\theta}$). How to determine if this estimate is good or bad ?

Bias $= E_\theta(\hat{\theta}) - \theta$ Standard Error : $\sqrt{Var_\theta(\hat{\theta})}$

MSE $= E_\theta((\theta - \hat{\theta})^2)$

Bias measures how close the average value of $\hat{\theta}$ to the true parameter θ

SE measures how variable is $\hat{\theta}$ around this average value

$$MSE = E_\theta((\theta - \hat{\theta})^2)$$
$$= E_\theta(\theta^2 - 2\theta\hat{\theta} + \hat{\theta}^2)$$
$$= E(\theta^2) - 2E(\theta)E(\hat{\theta}) + E(\hat{\theta}^2)$$

$= (\mathbb{E}(\hat{\theta}^2) - 2\mathbb{E}(\hat{\theta})\mathbb{E}(\hat{\theta}) + \mathbb{E}(\hat{\theta})^2) + \mathrm{Var}(\hat{\theta})$

$= \mathrm{Var}(\hat{\theta}) + (\mathbb{E}(\hat{\theta}) - \theta)^2$

$= \mathrm{Var}(\hat{\theta}) + \mathrm{Bias}^2$

* Consistency : An estimator $\hat{\theta}$ based on $X_1, \dots, X_n$ is consistent if $\hat{\theta} \xrightarrow{\mathbb{P}} \theta$ as $n \to +\infty$ ($\mathbb{P}(|\hat{\theta} - \theta| < \varepsilon) \longrightarrow 1 \ \forall \varepsilon$)

Working conditions : $\mathbb{E}(\hat{\theta}) = \theta$

$\mathrm{Var}(\hat{\theta}) \xrightarrow{n \to +\infty} 0$

* Examples :

1) $\hat{\lambda} = \bar{X}$ by MOM and MLE , $X_1, \dots, X_n \sim \mathcal{P}(\lambda)$

$\mathbb{E}(\hat{\lambda}) = \mathbb{E}(\bar{X}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(X_i) = \lambda$ Unbiased

$\mathrm{Var}(\hat{\lambda}) = \mathrm{Var}(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{\lambda}{n} \xrightarrow[\text{consistent}]{n \to +\infty} 0$, $\mathrm{Sd}(\hat{\lambda}) = \sqrt{\frac{\lambda}{n}}$

$\mathrm{MSE} = \mathrm{Bias}^2 + \mathrm{Var}(\hat{\lambda}) = 0 + \frac{\lambda}{n} = \frac{\lambda}{n}$

2) $\hat{\lambda} = \frac{1}{\bar{X}}$ for MOM and MLE . $X_1, \dots, X_n \sim \mathrm{Exp}(\lambda)$

Bias : $\mathbb{E}(\hat{\lambda}) = \mathbb{E}\left(\frac{n}{\sum X_i}\right) = \frac{n\lambda}{n-1}$ $\bar{X} \sim \mathrm{Gamma}(n, n\lambda)$

Jensen's Inequality : $\varphi(\mathbb{E}(\bar{X})) \leq \mathbb{E}(\varphi(\bar{X}))$

$\lambda = \frac{1}{\mathbb{E}(\bar{X})} \leq \mathbb{E}\left(\frac{1}{\bar{X}}\right) = \mathbb{E}(\hat{\lambda})$ so it is unbiased

3) Nonsense Unbiased Estimator

Let $X \sim \mathcal{P}(\lambda)$, $\lambda > 0$. We want to estimate the parameter $\theta = e^{-3\lambda}$

based on a sample of size 1. $T(X) = (-2)^X$

$$\mathbb{E}(T(X)) = \mathbb{E}((-2)^X) = \sum_{k\geq 0} \frac{(-2)^k e^{-\lambda} \cdot \lambda^k}{k!}$$

$$= e^{-\lambda} \sum_{k\geq 0} \frac{(-2\lambda)^k}{k!}$$

$$= e^{-\lambda} \cdot e^{-2\lambda}$$

$$= e^{-3\lambda}$$

Hence, this is not a good estimator as when x is odd the estimate is negative but λ is a positive quantity.

* **Remark** : When n is large, asymptotic theory provides us with a more complete picture of the "accuracy" of λ. By CLT $\sqrt{n}(\bar{X} - \lambda) \xrightarrow{a} N(0, \lambda)$ as $n \to +\infty$
a : asymptotically

* Asymptotically Normal : $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \theta)$

Finding Asymptotic Variance :

$$f(n|\lambda) = \frac{e^{-\lambda} \lambda^n}{n!} \qquad \log(f(n|\lambda)) = -\lambda + n\log(\lambda) - \log(n!)$$

$$z(n, \lambda) = \frac{\partial}{\partial \lambda} (\log(f(n|\lambda)) = \frac{n}{\lambda} - 1$$

$$z'(n, \lambda) = \frac{-n}{\lambda^2}$$

Fisher Information : $I(\lambda) = -\mathbb{E}_\lambda(z'(X, \lambda)) = -\mathbb{E}\left(\frac{-X}{\lambda^2}\right) = \frac{1}{\lambda}$

Hence $\hat{\theta} \xrightarrow{a} N\left(\theta, \frac{1}{n I(\theta)}\right) \Rightarrow \sqrt{n}(\hat{\theta} - \theta) \xrightarrow{a} N(\theta, \theta)$

Hence, $\hat{\theta} \sim N\left(\theta_0, \frac{1}{n\, I(\theta_0)}\right) \Leftrightarrow \sqrt{n}(\hat{\theta}-\theta_0) \sim N\left(0, \frac{1}{I(\theta_0)}\right)$

* **Theorem** :

Let $\{f(x|\theta) \mid \theta \in \Omega\}$ be a parametric model $\theta \in \mathbb{R}$ is a single parameter. Let $X_1, \ldots, X_n \overset{iid}{\sim} f(x|\theta_0)$ for $\theta_0 \in \Omega$. Let $\hat{\theta}$ be the MLE based on $X_1, \ldots, X_n$. Suppose certain regularity conditions hold.

1) All the pdf / pmf $f(x|\theta)$ in the model have the same support.
2) θ_0 is an interior point (not on the boundary)
3) The log likelihood $\ell(\theta)$ is differentiable w.r.t θ
4) $\hat{\theta}$ is the unique value (MLE) of $\theta \in \Omega$ that solves $\ell'(\theta) = 0$

Then, $\hat{\theta}$ is consistent and asymptotically normal, with

$$\sqrt{n}(\hat{\theta}-\theta_0) \xrightarrow{d} N\left(0, \frac{1}{I(\theta_0)}\right)$$

$I(\theta) := V_\theta(z(X,\theta)) = -E(z'(X,\theta))$ where $z(X,\theta)$ is the score function.

* **Proof** :

$$E(z(X,\theta)) = \int z(x,\theta)\, f(x|\theta)\, dx$$

$$= \int \frac{\partial}{\partial \theta} \log(f(x|\theta))\, f(x|\theta)\, dx$$

$$= \int \frac{\partial/\partial\theta\, f(x|\theta)}{f(x|\theta)} \cdot f(x|\theta)\, dx$$

$$= \int \frac{\partial}{\partial \theta} f(x|\theta)\, dx$$

$$= \frac{\partial}{\partial \theta} \int f(x|\theta)\, dx \quad 1$$

$$= 0$$

$$\mathbb{E}(z'(x,\theta)) = \int z'(x,\theta)\, f(x|\theta)\, dx$$

$$= \int \frac{\partial^2}{\partial^2 \theta} \log(f(x|\theta))\, f(x|\theta)\, dx$$

$$= \int \frac{\partial}{\partial \theta} z(x,\theta)\, f(x|\theta)\, dx$$

$$= \frac{\partial}{\partial \theta} \int z(x,\theta)\, f(x|\theta)\, dx$$

$$= \frac{\partial}{\partial \theta} \mathbb{E}(z(x,\theta))$$

$$\mathbb{E}(z'(x,\theta)) = \frac{\partial}{\partial \theta} \mathbb{E}(z(x,\theta))$$

$$0 = \frac{\partial}{\partial \theta} \int z(x,\theta)\, f(x|\theta)\, dx$$

$$= \int \frac{\partial}{\partial \theta} (z(x,\theta)\, f(x|\theta))\, dx$$

$z(x,\theta) = \frac{\partial}{\partial \theta} \log(f(x|\theta)) = \frac{\partial/\partial\theta\, f(x|\theta)}{f(x|\theta)}$

$$= \int z'(x,\theta) f(x|\theta) + z(x,\theta) \frac{\partial}{\partial \theta} f(x|\theta) \quad = \quad z(x,\theta) f(x|\theta)$$

$$= \int z'(x,\theta) f(x|\theta)\, dx + \int z^2(x,\theta) f(x|\theta)\, dx$$

$$= \mathbb{E}(z'(x,\theta)) + \mathbb{E}(z^2(x,\theta))$$

$$= \mathbb{E}(z'(x,\theta)) + \mathrm{Var}(z(x,\theta)) \quad \text{as } \mathbb{E}(z(x,\theta))$$

Hence, $\mathrm{Var}(z(x,\theta)) = -\mathbb{E}(z'(x,\theta))$

* Proof sketch : Assuming $f(x|\theta)$ is the PDF

To see why $\hat{\theta}_{MLE}$ is consistent, we note that $\hat{\theta}$ is the value of θ that maximizes :

$$\frac{1}{n} l(\theta) = \frac{1}{n} \log\left(\prod_{i=1}^{n} f(x|\theta)\right) = \frac{1}{n} \sum_{i=1}^{n} \log(f(x|\theta))$$

Suppose $X_1, \dots, X_n \overset{iid}{\sim} f(x|\theta_0)$. By LLN

$$\frac{1}{n} \sum_{i=1}^{n} \log(f(x|\theta)) \xrightarrow{P} E(\log(f(x|\theta)))$$

$$E(\log(f(x|\theta)) - E(\log(f(x|\theta_0)) = E\left(\log\left(\frac{f(x|\theta)}{f(x|\theta_0)}\right)\right)$$

By Jensen's Inequality ($x \mapsto \log$ is concave):

$$E\left(\log\left(\frac{f(x|\theta)}{f(x|\theta_0)}\right)\right) \leq \log\left(E\left(\frac{f(x|\theta)}{f(x|\theta_0)}\right)\right)$$

$$= \log\left(\int \frac{f(x|\theta)}{f(x|\theta_0)} f(x|\theta_0)\, dx\right)$$

$$= \log\left(\int f(x|\theta)\, dx\right)$$

$$= \log(1)$$

$$= 0$$

Hence, θ is maximized at θ_0

By the definition of MLE : $l'(\hat{\theta}) = 0$

Consistency of $\hat{\theta}$ ensures that $\hat{\theta}$ is close to the true value of θ i.e θ_0 with high probability.

$$l'(\theta_0) + (\hat{\theta} - \theta_0)\, l''(\theta_0) = 0 \quad \text{by Taylor's Expansion}$$

$$\hat{\theta} - \theta_0 = - \frac{l'(\theta_0)}{l''(\theta_0)}$$

$$\sqrt{n}\,(\hat{\theta} - \theta) = - \frac{\frac{1}{\sqrt{n}}\, l'(\theta_0)}{\frac{1}{n}\, l''(\theta_0)}$$

$$\frac{1}{n}\, l''(\theta) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log(f(x_i|\theta))$$

$$= \frac{1}{n} \sum_{i=1}^{n} z'(x_i, \theta)$$

$$\xrightarrow{\mathbb{P}} \mathbb{E}(z'(X, \theta)) = - I(\theta)$$

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log(f(x_i|\theta)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} z(x_i, \theta)$$

$$\longrightarrow N(0, I(\theta))$$

By Slutsky's Lemma: $\sqrt{n}\,(\hat{\theta} - \theta) \sim \frac{1}{I(\theta)} N(0, I(\theta))$

$$\sqrt{n}\,(\hat{\theta} - \theta) \sim N\left(0, \frac{1}{I(\theta)}\right)$$

* **Remark**: 1) $\hat{\theta}$ is asymptotically unbiased. More precisely the bias of $\hat{\theta}$ is less than the order $\frac{1}{\sqrt{n}}$. Otherwise $\sqrt{n}\,(\hat{\theta} - \theta_0)$ does not converge to a zero mean distribution.

2) The variance of $\hat{\theta}$ is approximately $\frac{1}{n I(\theta)}$.
In particular, the standard error is of order $\frac{1}{\sqrt{n}}$ and the variance is the main contributor to the MSE = Var + Bias² of $\hat{\theta}$.

3) If the true parameter is θ_0, then the sampling distribution of $\hat{\theta}$ is approximately $N\left(\theta_0, \frac{1}{n I(\theta)}\right)$

* **Example:** Sampling Distribution of Gamma

$X_1, \dots, X_n \overset{iid}{\sim} \text{Gamma}(\alpha, 1)$

$$\log(f(x|\alpha)) = \log\left(\frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x}\right)$$

$$= (\alpha-1)\log x - x - \log(\Gamma(\alpha))$$

$$z(x,\alpha) = \log(x) - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$$

$$z'(x,\alpha) = -\Psi'(\alpha), \quad \Psi(\alpha) = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$$

$$I(\alpha) = -\mathbb{E}(z'(X,\alpha)) = \Psi'(\alpha)$$

$$\sqrt{n}(\hat{\alpha}_{MLE} - \alpha) \sim N\left(0, \frac{1}{\Psi'(\alpha)}\right)$$

$$\hat{\alpha}_{MLE} \sim N\left(\alpha, \frac{1}{n\Psi'(\alpha)}\right)$$

* **Fisher Information for more that one parameter:**

Let $X_1, \dots, X_n \overset{iid}{\sim} f(x|\theta)$, $\{f(x|\theta) \mid \theta \in \mathbb{R}^k\}$ with k

parameters and $\hat{\theta}_n$ is the MLE. We define F. I matrix $I(\theta) \in \mathbb{R}^{k \times k}$ is given by:

$$I_{(i,j)} = Cov\left(\frac{\partial}{\partial \theta_i} \log(f(X|\theta)), \frac{\partial}{\partial \theta_j} \log(f(X|\theta))\right)$$

$$= -\mathbb{E}\left(\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log(f(X|\theta))\right)$$

$\sqrt{n}(\hat{\theta}_n - \theta) \sim MVN(0, I(\theta)^{-1})$ where $I(\theta)^{-1}$ is the $k \times k$ matrix inverse of $I(\theta)$.

* **Example:** Sampling Distribution of Gamma

$X_1, \ldots, X_n \overset{iid}{\sim} Gamma(\alpha, \beta)$

$$\log(f(x|\alpha,\beta)) = \log\left(\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}\right)$$

$$= \alpha \log(\beta) + (\alpha - 1)\log x - \beta x - \log(\Gamma(\alpha))$$

$$= S$$

$$\frac{\partial^2}{\partial \alpha^2} S = -\Psi'(\alpha)$$

$$\frac{\partial^2}{\partial \beta^2} S = \frac{\partial}{\partial \beta}\left(\frac{\alpha}{\beta} - x\right) = -\frac{\alpha}{\beta^2}$$

$$\frac{\partial^2 S}{\partial \alpha \partial \beta} = \frac{\partial}{\partial \alpha}\left(\frac{\alpha}{\beta} - x\right) = \frac{1}{\beta}$$

Hence, $I(\alpha, \beta) = \begin{pmatrix} \Psi'(\alpha) & -1/\beta \\ -1/\beta & \alpha/\beta^2 \end{pmatrix}$

$$I^{-1}(\alpha, \beta) = \frac{\beta^2}{\alpha\Psi'(\alpha) - 1} \begin{pmatrix} \alpha/\beta^2 & 1/\beta \\ 1/\beta & \Psi'(\alpha) \end{pmatrix}$$

$$(\hat{\alpha}, \hat{\beta}) \sim BVN\left((\alpha, \beta), \frac{1}{n} I^{-1}(\alpha, \beta)\right)$$

$$\hat{\alpha} \sim N\left(\alpha, \frac{\beta^2}{\alpha\Psi'(\alpha) - 1} \cdot \frac{\alpha}{n\beta^2}\right)$$

$$\hat{\alpha} \sim N\left(\alpha, \frac{\alpha}{n(\alpha\Psi'(\alpha) - 1)}\right)$$

Hence, larger model has higher variability.

Implication: A complex model may better capture the true distribution of the data but they will be increasing the difficulty in estimating the parameters than those in a simple model.

Q_1 : Why is Fisher information $I(\theta)$ called information?

Q_2 : Why should we use MLE for estimating θ?

Let $X_1, \ldots, X_n \overset{iid}{\sim} f(x|\theta_0)$, $l(\theta) = \sum_{i=1}^{n} \log(f(x_i|\theta))$ and $I(\theta_0) = -E\left(\frac{\partial^2}{\partial\theta^2} \log(f(X_i|\theta))|_{\theta=\theta_0}\right) = -\frac{1}{n} E(l''(\theta_0))$. Fisher information matrix measures the expected curvature of the log likelihood function $l(\theta)$ around the true parameter $\theta = \theta_0$.

A_1: F.I quantifies the amount of information that each observation X_i contains about the unknown parameter.

* **Theorem**: Cramer Rao Lower Bound (CRLB)

Consider a parametric model $\{f(x|\theta) \mid \theta \in \Omega\}$ satisfying the regularity conditions where $\theta \in \mathbb{R}$. Let T be an unbiased estimator of θ based on the data $X_1, \ldots, X_n \overset{iid}{\sim} f(x|\theta)$. Then, $\text{Var}(T) \geq \frac{1}{nI(\theta)}$

A_2: The CRLB ensures that no unbiased estimator can achieve asymptotically lower variance than the MLE. In general, MOM estimator have higher MSE than MLE for large n.

* **Efficiency**: Let T_1 and T_2 be two unbiased estimators of θ.

Relative efficiency $= \frac{\text{Var}(T_1)}{\text{Var}(T_2)}$.

An unbiased estimator is efficiency if its variance is $= \frac{1}{nI(\theta)}$. We say MLE is asymptotically efficiency.

* **MLE under model misspecification**:

If the (parametric) model is incorrect, does MLE estimate $\hat{\theta}$ still remain meaningful?

We have been measuring the error of an estimator $\hat{\theta}$ by bias, variance and MSE.

* **Def**: KL Divergence

If $X_1, \ldots, X_n \overset{iid}{\sim} g$ that is not in the model (parametric family $\{f(x|\theta) \mid \theta \in \Omega\}$) then there is no true parameter θ associated to g.

Now, we introduce a measure of "distance" between two PDFs.

For two PDFs f and g, the KL divergence from f to g is given by:

$$D_{KL}(g \| f) = \int g(x) \log\left(\frac{g(x)}{f(x)}\right) dx.$$

Equivalently, if $X \sim g$, then $D_{KL}(g \| f) = \mathbb{E}\left(\log\left(\frac{g(x)}{f(x)}\right)\right)$

1) $f = g$ then $D_{KL}(g \| f) = 0$

2) $\mathbb{E}\left(\log\left(\frac{g(x)}{f(x)}\right)\right) = \mathbb{E}\left(-\log\left(\frac{f(x)}{g(x)}\right)\right)$

$$\geq -\log\left(\mathbb{E}\left(\frac{f(x)}{g(x)}\right)\right)$$

$$\geq -\log\left(\int \frac{f(x)}{g(x)} \cdot g(x)\, dx\right)$$

$$\geq -\log(1)$$

$$\geq 0$$

3) If g and f are constants then $D_{KL}(g \| f) = 0$

* Example :

Suppose $f = N(\mu_0, \sigma^2)$ and $g = N(\mu_1, \sigma^2)$. Calculating $D_{KL}(g \| f)$ if $X \sim g$ and $D_{KL}(f \| g)$.

$$D_{KL}(g \| f) = \mathbb{E}\left(\log\left(\frac{g(X)}{f(X)}\right)\right)$$

$$\log\left(\frac{g(x)}{f(x)}\right) = \log\left(\frac{e^{-\frac{(x-\mu_1)^2}{2\sigma^2}}}{e^{-\frac{(x-\mu_0)^2}{2\sigma^2}}}\right) = \frac{-(x-\mu_1)^2}{2\sigma^2} + \frac{(x-\mu_0)^2}{2\sigma^2}$$

$$= \frac{-x^2 + 2x\mu_1 - \mu_1^2 + x^2 - 2x\mu_0 + \mu_0^2}{2\sigma^2}$$

$$= \frac{(\mu_0^2 - \mu_1^2) + 2x(\mu_1 - \mu_0)}{2\sigma^2}$$

$$\mathbb{E}\left(\log\left(\frac{g(X)}{f(X)}\right)\right) = \mathbb{E}\left(\frac{(\mu_0^2 - \mu_1^2) + 2x(\mu_1 - \mu_0)}{2\sigma^2}\right)$$

$$= \frac{(\mu_0^2 - \mu_1^2)}{2\sigma^2} + \frac{(\mu_1 - \mu_0)}{\sigma^2} \cdot \mu_1$$

$$= \frac{\mu_0^2 - \mu_1^2 + 2\mu_1^2 - 2\mu_0\mu_1}{2\sigma^2}$$

$$= \frac{\mu_0^2 - 2\mu_0\mu_1 + \mu_1^2}{2\sigma^2}$$

$$= \frac{(\mu_0 - \mu_1)^2}{2\sigma^2}$$

$$D_{KL}(f \| g) = \mathbb{E}\left(\log\left(\frac{f(X)}{g(X)}\right)\right)$$

$$\log\left(\frac{f(x)}{g(x)}\right) = \log\left(\frac{e^{-\frac{(x-\mu_0)^2}{2\sigma^2}}}{e^{-\frac{(x-\mu_1)^2}{2\sigma^2}}}\right) = \frac{-(x-\mu_0)^2}{2\sigma^2} + \frac{(x-\mu_1)^2}{2\sigma^2}$$

$$= \frac{-x^2 + 2x\mu_0 - \mu_0^2 + x^2 - 2x\mu_1 + \mu_1^2}{2\sigma^2}$$

$$= \frac{(\mu_1^2 - \mu_0^2) + 2x(\mu_0 - \mu_1)}{2\sigma^2}$$

$$\mathbb{E}\left(\log\left(\frac{f(X)}{g(X)}\right)\right) = \mathbb{E}\left(\frac{(\mu_1^2 - \mu_0^2) + 2x(\mu_0 - \mu_1)}{2\sigma^2}\right)$$

$$= \frac{(\mu_1^2 - \mu_0^2)}{2\sigma^2} + \frac{(\mu_0 - \mu_1)}{\sigma^2} \cdot \mu_0$$

$$= \frac{\mu_1^2 - \mu_0^2 + 2\mu_0^2 - 2\mu_0\mu_1}{2\sigma^2}$$

$$= \frac{\mu_0^2 - 2\mu_0\mu_1 + \mu_1^2}{2\sigma^2}$$

$$= \frac{(\mu_0 - \mu_1)^2}{2\sigma^2}$$

* Let $X_1, X_2, \dots, X_n \sim g$, we suppose that $D_{KL}(g \| f)$ has a unique minimum $\theta = \theta^*$. Then, under suitable regularity conditions, the MLE $\hat{\theta}$ converges to θ^* in probability as $n \to +\infty$. Then, we call $f(x|\theta^*)$ as the KL projection of g onto the parametric model $\{f(x|\theta) \mid \theta \in \Omega\}$. In other words, the MLE is estimating the distribution in our model that is closest w.r.t KL divergence to g.

* **Sandwich estimator of variance** :

Let $X_1, \dots, X_n \overset{iid}{\sim} g$. how close is the MLE $\hat{\theta}$ to this KL-projection.

$0 = \ell'(\hat{\theta})$, we write Taylor expansion around $\hat{\theta} = \theta^*$

$0 \approx \ell'(\theta^*) + (\hat{\theta} - \theta^*)\ell''(\theta^*)$

$$\Rightarrow (\hat{\theta} - \theta^*) = -\frac{\ell'(\theta^*)}{\ell''(\theta^*)}$$

$$\Rightarrow \sqrt{n}(\hat{\theta} - \theta^*) = \frac{-\frac{1}{\sqrt{n}}\ell'(\theta^*)}{\frac{1}{n}\ell''(\theta^*)} = \frac{\frac{1}{\sqrt{n}}\ell'(\theta^*)}{-\mathbb{E}(z'(x,\theta^*))} \xrightarrow{\text{asymptotically } z(x,\theta) \to N(0, Var(z(x,\theta^*)))} N\left(0, \frac{Var(z(x,\theta^*))}{\mathbb{E}(z'(x,\theta^*)^2}\right)$$

* **Example** :

$X_1, \dots, X_n \overset{iid}{\sim} Exp(\lambda)$

$$0 = \frac{\partial}{\partial\lambda}\left(\log\left(\lambda^n e^{-\lambda\sum_{i=1}^n x_i}\right)\right) = \frac{\partial}{\partial\lambda}\left(n\log\lambda - \lambda\sum_{i=1}^n x_i\right) = \frac{n}{\lambda} - \sum_{i=1}^n x_i$$

$$\hat{\lambda}_{MLE} = \frac{1}{\bar{x}}$$

$$z(x,\lambda) = \frac{\partial}{\partial\lambda}(\log(f(x|\lambda)) = \frac{\partial}{\partial\lambda}(\log\lambda - \lambda x) = \frac{1}{\lambda} - x$$

$$z'(x,\lambda) = \frac{\partial}{\partial\lambda} z'(x,\lambda) = \frac{-1}{\lambda^2}$$

$$\bar{z}(x_i,\lambda) = \frac{1}{\lambda} - \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{\lambda} - \bar{x}$$

$$\mathbb{E}(z'(x,\theta^*)) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{-1}{\lambda^2}\right) = \frac{-1}{\lambda^2} = -\bar{x}^2$$

$$Var(z(x,\theta^*)) = \frac{1}{n-1}\sum_{i=1}^{n}(z(x_i,\theta^*) - \bar{z}(x_i,\theta^*))^2$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{1}{\lambda} - x - \frac{1}{\lambda} + \bar{x}\right)^2$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}(x-\bar{x})^2$$

$$= s_x^2$$

Sandwich estimate of Variance : $\frac{Var(z(x,\theta^*))}{\mathbb{E}(z(x,\theta^*))^2} = \frac{s_x^2}{\bar{x}^4}$

* Plugin Estimater :

So far, we know how to estimate θ. Then, we learnt how to compute MLE. Then, we learnt the sampling distribution and variance for MLE. No unbiased estimater achieve smaller than that of the MLE for large n (CRLB).

Now, we are interested in estimating $g(\theta)$ by $g(\hat{\theta})$, where $\hat{\theta}$ (say MLE) is an estimate of θ. This is called plugin estimater.

* Example : Pareto Distribution

$f(x|\theta, x_0) = \theta x_0^{\theta} x^{-\theta-1} \mathbb{1}(x \geq x_0)$, where $\theta > 1$

$f(x|\alpha, x_0) = \alpha \cdot x_0^{\alpha} \cdot x^{-\alpha-1} \mathbb{1}(x \geq x_0)$ where $\alpha > 1$

Used as an income distribution and x_0 represents the minimum possible income.

Barabasi Network Talk

One parameter pareto distribution : $f(x) = \alpha \cdot x^{-\alpha-1} \mathbb{1}(x \geq 1)$

$$\mathbb{E}(X) = \int_{\mathbb{R}} x f(x|\alpha)\, dx$$

$$= \int_1^{+\infty} \alpha \cdot x^{-\alpha}\, dx$$

$$= \alpha \lim_{n \to +\infty} \left[\frac{x^{-\alpha+1}}{-\alpha+1} \right]_1^n$$

$$= \alpha \lim_{n \to +\infty} \left(\frac{n^{1-\alpha}}{1-\alpha} + \frac{1}{\alpha-1} \right) \qquad 1-\alpha < 0 \text{ as } \alpha > 1$$

$$= \frac{\alpha}{\alpha-1}$$

$$\ell(\alpha) = n \log(\alpha) - (\alpha+1) \sum_{i=1}^{n} \log(x_i)$$

$$0 = \frac{n}{\alpha} - \sum_{i=1}^{n} \log(x_i)$$

$$\hat{\alpha}_{MLE} = \frac{n}{\sum_{i=1}^{n} \log(x_i)}$$

Hence, the average income is estimated by : $\dfrac{\hat{\alpha}_{MLE}}{\hat{\alpha}_{MLE} - 1}$

We we don't have enough data but we know the distribution of the data the we can still estimate $g(\alpha)$.

* **Examples** : Odds (Odds Ratio)

You play a game with a friend, where you flip a biased coin. If

the coin lands heads, then you give your friend 1 AED. If the coin lands tails, your friend gives you n AED. What is the value of n that makes this a fair coin?

Fair game $\Rightarrow$ earning = 0 AED

$$E(\text{winning}) = -p + (1-p)n = 0 \Rightarrow n = \frac{p}{1-p}$$ odds of getting heads over tails

$X \sim Ber(p)$

$$\ell(p) = \sum_{i=1}^{n} n_i \log(p) + \sum_{i=1}^{n} (1-n_i) \log(1-p)$$

$$\ell'(p) = \frac{\sum n_i}{p} - \frac{n - \sum n_i}{1-p} = 0$$

$$(1-p)\sum_{i=1}^{n} n_i - np + p\sum_{i=1}^{n} n_i = 0$$

$$\sum_{i=1}^{n} n_i - np = 0$$

$$\hat{p}_{MLE} = \bar{x}$$

Hence, the estimate $\frac{\bar{x}}{1-\bar{x}}$

* **Delta Method**:

$$g(\hat{\theta}) = g(\theta_0) + (\hat{\theta} - \theta_0) g'(\theta_0)$$ Taylor expansion around θ_0

$$g(\hat{\theta}) - g(\theta_0) = (\hat{\theta} - \theta_0) g'(\theta_0)$$

$$\Rightarrow \sqrt{n}\,(g(\hat{\theta}) - g(\theta_0)) = \sqrt{n}\,(\hat{\theta} - \theta_0) g'(\theta_0) \sim N\left(0, \frac{g'(\theta_0)^2}{I(\theta_0)}\right)$$

* Using the delta method, we can quantify the uncertainty about the plugin estimator $g(\hat{\theta})$ for large n.

* Restrictions : 1) $g : \mathbb{R} \to \mathbb{R}$ real valued

2) $g'(\theta_0) \neq 0$

$$\sqrt{n}\,(g(\hat{\theta}) - g(\theta_0)) \underset{\text{in distribution}}{\sim} N\left(0, \frac{g'(\theta_0)^2}{I(\theta)}\right) \quad \text{as } n \to +\infty$$

* **Example** : Log-odds

Let $X_1, X_2, \dots, X_n \overset{iid}{\sim} Ber(p)$ and recall the estimate of log-odds $\log\left(\frac{p}{1-p}\right)$ given by $\log\left(\frac{\bar{X}}{1-\bar{X}}\right)$.

By CLT : $\sqrt{n}\,(\bar{X} - p) \sim N(0, p(1-p))$

$$(\log(p) - \log(1-p))' = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}$$

By the delta method : $\sqrt{n}\left(\log\left(\frac{\bar{X}}{1-\bar{X}}\right) - \log\left(\frac{p}{1-p}\right)\right) \sim N\left(0, \frac{1}{p(1-p)}\right)$

Interpretation : Our estimate log-odds of heads to tails is approximately normally distributed around the true odds $\log\left(\frac{p}{1-p}\right)$ with variance $\frac{1}{np(1-p)}$.

Suppose we toss a coin 1000 times. We observe 600 heads.

$\bar{X} = \hat{p} = \frac{600}{1000} = 0.6$, we estimate log-odds by $\log\left(\frac{0.6}{0.4}\right) = 0.41$ and standard error $\sqrt{\frac{1}{n\bar{X}(1-\bar{X})}} \approx 0.2$.

* **Example** : Pareto Distribution

We showed that the average income is estimated by $g(\theta) = \frac{\theta}{\theta - 1}$ where $\hat{\theta}_{MLE} = \frac{n}{\sum \log(x_i / x_m)}$. Also, $g'(\theta) = \frac{-1}{(\theta-1)^2}$.

$\sum_{i=1}^{n} \log(x_i)$ $\qquad\qquad$ $(\theta-1)^2$

Hence, by the delta method : $\sqrt{n}\left(\frac{\hat{\theta}}{\theta-1} - \frac{\theta}{\theta-1}\right) \sim N\left(0, \frac{\theta^2}{(\theta-1)^4}\right)$

$$
\begin{aligned}
I(\theta) &= -E(l'(X, \theta)) \\
&= -E\left(\frac{\partial^2}{\partial \theta}\left(\log(\theta) - (\theta+1)\log(x)\right)\right) \\
&= -E\left(\frac{\partial}{\partial \theta}\left(\frac{1}{\theta} - \log(x)\right)\right) \\
&= -E\left(\frac{-1}{\theta^2}\right) \\
&= \frac{1}{\theta^2}
\end{aligned}
$$

Hence, $\sqrt{n}(\hat{\theta} - \theta) \sim N(0, \theta^2)$.

Finally, $\frac{\hat{\theta}}{\hat{\theta}-1} \sim N\left(\frac{\theta}{\theta-1}, \frac{\theta^2}{n(\theta-1)^4}\right)$.

* **Theorem** : General Cramer Rao Lower Bound

For a parametric model $\{f(x|\theta) \mid \theta \in \Omega\}$ satisfying certain mild regularity conditions and g be any function differentiable on all of Ω and T be any unbiased estimator of $g(\theta)$ based on data $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f(x|\theta)$. Then, $V(T) \geq \frac{g'(\theta)^2}{n\, I(\theta)}$.

When an estimator T achieves this variance, it is called efficient.

* **Properties of Estimators** :

From what we learnt consistency and asymptotic efficiency are large sample property whereas unbiasedness and minimum variance are practical properties. Are there any more practical properties ?

* Def : Minimum Variance Unbiased Estimator (MVUE)

Unbiased estimators can average hit the target for all the parameter values. This seems to be a reasonable constraint to impose on an estimator and indeed produces meaningful estimates in a wide variety of situations. In a large class of problems, it is possible to find an UE of $g(\theta)$ that has the smallest possible variance among all these unbiased estimators. Such an estimate is called MVUE

1) $E_\theta(S_n) = g(\theta)$
2) $Var_\theta(S_n) \leq V_\theta(T_n)$ where T_n is an unbiased estimator of $g(\theta)$

* Example : Normal Distribution

$X_1, \dots, X_n \sim N(\mu, \sigma^2)$

$\bar{X}$: sample mean

$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$

$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$

$E(\bar{X}) = \frac{1}{n} \sum_{i=1}^{n} E(X_i) = \mu$ unbiased

$Var(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^{n} Var(X_i) = \frac{\sigma^2}{n} \longrightarrow 0$

Hence, $\bar{X}$ is consistant & MVUB as it achieves the lowest variance.

$E(\hat{\sigma}^2) = E\left(\frac{n-1}{n} S^2\right) = \frac{(n-1)\sigma^2}{n}$

$E(S^2) = \frac{1}{n-1} \sum_{i=1}^{n} E((X_i - \bar{X})^2) = \frac{1}{n-1} \sum_{i=1}^{n} E((X_i - \mu + \mu - \bar{X})^2)$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n} E((x_i-\mu)^2) + \sum_{i=1}^{n} E((\mu-\bar{x})^2) - 2\sum_{i=1}^{n} E((x_i-\mu)(\mu-\bar{x})\right)$$

$$= \frac{1}{n-1}\left(n\,Var(x_i) + n\,Var(\bar{x}) - 2n\,E((\mu-\bar{x}^2))\right)$$

$$= \frac{1}{n-1}\left(n\sigma^2 + \frac{\sigma^2}{n}\cdot n - 2n\,Var(\bar{x})\right)$$

$$= \frac{1}{n-1}\left(\sigma^2(n-1)\right)$$

$$= \sigma^2 \quad \text{MVUB}$$

Unbiased estimates may not always be better than biased estimators. In ML, performance vs model complexity trade-off

* **Example** : $x_1, x_2, \dots, x_n \overset{iid}{\sim} U(0,\theta)$, $\theta>0$ $\mathcal{X} =]0,+\infty[$. A natural estimate of θ is $x_{(n)} = \max\{x_i\}$

$E(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n} E(\theta_i) = \frac{\theta}{2}$ so $2\bar{x}$ is an unbiased estimator of θ.

Show that $x_{(n)}$ outperforms $2\bar{x}$ by an order of magnitude in the sense of MSE. Then, find out the MVUE of θ.

- $E(2\bar{x}) = \frac{2}{n}\sum_{i=1}^{n} E(x_i) = \theta$

- $Var(2\bar{x}) = \frac{4}{n^2}\sum_{i=1}^{n} Var(x_i)$

$$= \frac{4}{n}\,Var(x_i) \qquad E(x_i^2) = \int_0^\theta \frac{u^2}{\theta}\,du = \frac{u^3}{3\theta}\Big|_0^\theta = \frac{\theta^2}{3}$$

$$= \frac{4}{n}\left(\frac{\theta^2}{3} - \frac{\theta^2}{4}\right)$$

$= \frac{\theta^2}{3n}$

- $MSE(2\bar{X}, \theta) = Var(2\bar{X}) + Bias^2 = \frac{\theta^2}{3n}$

- $F_n(x) = \mathbb{P}(X_{(n)} \leq x)$
  $= (F_X(x))^n$

  $f_n(x) = n\,(F_X(x))^{n-1} \cdot f_X(x)$
  $= n \left(\frac{x}{\theta}\right)^{n-1} \cdot \frac{1}{\theta}\, \mathbb{1}_{]0,\theta[}$

  $f_n(x) = \frac{n x^{n-1}}{\theta^n}\, \mathbb{1}_{]0,\theta[}(x)$

- $\mathbb{E}(X_{(n)}) = \frac{n}{\theta^n} \int_0^\theta x^n \, dx$
  $= \frac{n}{\theta^n} \left(\frac{x^{n+1}}{n+1} \Big|_0^\theta \right)$
  $= \frac{n\theta}{n+1}$

  $\mathbb{E}(X_{(n)}^2) = \frac{n}{\theta^n} \int_0^\theta x^{n+1} \, dx$
  $= \frac{n}{\theta^n} \left(\frac{x^{n+2}}{n+2} \Big|_0^\theta \right)$
  $= \frac{n\theta^2}{n+2}$

- $Var(X_{(n)}) = \frac{n\theta^2}{n+2} - \frac{n^2\theta^2}{(n+1)^2} = \frac{n\theta^2}{(n+2)(n+1)^2}$

- $MSE(X_{(n)}, \theta) = \frac{n\theta^2}{(n+2)(n+1)^2} + \left(\frac{n\theta}{n+1} - \theta\right)^2$
  $= \frac{n^2\theta^2}{n(n+2)(n+1)^2} + \frac{\theta^2}{(n+1)^2}$

- $MSE\left(\left(1+\frac{1}{n}\right)X_{(n)}, \theta\right) = \frac{(n+1)^2}{n^2} \left(\frac{n^2\theta^2}{n(n+2)(n+1)^2}\right) = \frac{\theta^2}{n(n+2)}$

  Because $\mathbb{E}\left(\left(1+\frac{1}{n}\right)X_{(n)}\right) = \frac{n+1}{n}\left(\frac{n\theta}{n+1}\right) = \theta$ so it is unbiased.

* Sufficient Statistic :

A : Given $X_1, \dots, X_n$ B : Given $T = \phi(X_1, \dots, X_n)$


Generally, A will be able to find a better estimater than B. However, if B will be able to do just as good as A where $T = \phi(X_1, \dots, X_n)$ will in some sense summerize all the information contained in the ramdom sample about θ then knowledge of the whole data $X_1, \dots, X_n$ will be irrelevant in the search for a good and estimator of θ. A statistic T with such property is called a sufficient statistic.

* Def : Sufficient Statistic

Let $X_1, \dots, X_n$ be a ramdom sample from a distribution indexed by a parameter $\theta \in \Omega$. Let T be a statistic. Suppose that, for every $\theta \in \Omega$ and every possible value of T, the conditional distribution of $X_1, \dots, X_n$ given $T = t$ (at θ) depends only on t but not θ.

In conclusion, T is sufficient for obtaining as much info about θ as one could get from $(X_1, X_2, \dots, X_n)$.

If T is complete sufficient statistic, then $E(h(T)) = 0$

Lehmann Theorem : Achieves the minimum variance ⇒ UMVUE

* Example :

$X_1, \dots, X_n \sim P(\lambda)$, $\lambda > 0$ $\quad T = \sum_{i=1}^{n} X_i$

$$\mathbb{P}(\underset{\sim}{X} = \underset{\sim}{n} \mid T(X) = t) = \frac{\mathbb{P}(X = n, T(X) = t)}{\mathbb{P}(T(X) = t)}$$

$$= \frac{\mathbb{P}(X = n)}{\mathbb{P}(T(X) = t)} \quad \text{if } T(\underset{\sim}{n}) = t$$

$$= \frac{e^{-n\lambda} \lambda^{T(n)}}{\pi (n_i!)} \cdot \frac{t!}{e^{-n\lambda} \cdot (n\lambda)^t}$$

$$= \frac{t!}{\prod_{i=1}^{n} n_i! \, n^t} \quad \text{independent of } \lambda$$

Hence, it is a sufficient statistic

Trivial sufficient statistic : $(X_1, X_2, \ldots, X_n)$

$(X_{(1)}, X_{(2)}, \ldots, X_{(n)})$

* Finding a sufficient Statistic :

* **Theorem** : Neymen Fisher Factorization Theorem (TFFT)

Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f(n, \theta)$ where θ is unknown and $\theta \in \Omega$. A statistic $T(\underset{\sim}{n})$ is a sufficient statistic for θ iff the joint PDF / PMF $f_n(\underset{\sim}{n}, \theta)$ of $(X_1, \ldots, X_n)$ can be factorized as follows :

$$f_n(n, \theta) = u(\underset{\sim}{n}) \, v(T(\underset{\sim}{n}), \theta)$$

where u is independent of θ and v depends on θ and n through $T(n)$. Both u and v are non-negative.

* **Example** :

$X_1, \dots, X_n \sim P(\theta)$

$$f_n(n, \theta) = \prod_{i=1}^{n} \frac{e^{-\theta} \cdot \theta^{n_i}}{n_i!}$$

$$= \frac{e^{-n\theta} \cdot \theta^{\sum n_i}}{\left(\prod_{i=1}^{n} n_i!\right)}$$

$$= \frac{1}{\prod_{i=1}^{n} n_i!} \left(e^{-n\theta} \cdot \theta^{\sum n_i}\right)$$

$$= u(n) \cdot v(\theta, T(n)) \quad \text{where } T(n) = \sum_{i=1}^{n} n_i$$

* **Example** :

$$f_n(n|\theta) = \underbrace{\{\Gamma(\alpha)^n \cdot \left(\prod_{i=1}^{n} n_i\right)^{\alpha}\}}_{u(n)} \underbrace{\{\beta^{n\alpha} \cdot e^{-\frac{\beta}{c}\sum x_i}\}}_{v(\beta, \alpha, T)}$$

where $T(n) = \sum_{i=1}^{n} x_i$ where α is known and β unknown or $T(n) = \prod_{i=1}^{n} n_i$ where α is unknown and β known

* **Theorem** : Rao Blackwell Theorem

For every $\theta \in \Omega$, $MSE(\delta_0(T), g(\theta)) \leq MSE(\delta(X), g(\theta))$.

In summary Rao Blackwell Theorem shows how to improve upon an estimator that is not a function of a sufficient statistic by using a new estimator that is function of a sufficient statistic.

If we known a function $S_\theta(T) = \mathbb{E}_\theta(\delta(\underset{\sim}{X}) | T(\underset{\sim}{X}))$ where $T(\underset{\sim}{X})$ is a sufficient statistic and δ is an estimator of $g(\theta)$. The function S is indeed an estimator of $g(\theta)$ because it only depends on the observations X but not on θ.

* Confidence Intervals :

In a parametric model, let $g(\theta)$ be a quantity of interest. Informally, a confidence interval for $g(\theta)$ is a random interval calculated from the data that contains this value $g(\theta)$ with a specified probability. For example, 90% CI contains $g(\theta)$ with prob 0.9. It means if we conduct two different 90% CIs for $g(\theta)$ using 100 independent sets of data, then we expect about 90 of them to contain $g(\theta)$. "Idea of repeated sampling".

* Let $X_1, X_2, \dots, X_n$ be a data sample. By a random interval whose lower and upper endpoints $L(X_1, \dots, X_n)$ & $U(X_1, \dots, X_n)$ are random functions of $(X_1, \dots, X_n)$. The interval $[L(X_1, \dots, X_n), U(X_1, \dots, X_n)]$ is a $(1-\alpha)100\%$ confidence interval of $g(\theta)$ if $\mathbb{P}_\theta$ denotes the probability under $X_1, \dots, X_n \overset{iid}{\sim} f(x, \theta)$

$$\mathbb{P}_\theta\left(L(X_1, \dots, X_n) \le g(\theta) \le U(X_1, \dots, X_n)\right) = 1 - \alpha$$

* Example :

Let $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ where μ and σ^2 are unknown. Compute the CI of μ based on $\bar{X}$.

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{where} \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} \qquad \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \cdot \sqrt{\frac{\sigma^2}{S^2(n-1)}} \sim t_{n-1}$$

Let $t_{\alpha/2, n-1}$ and $-t_{\alpha/2, n-1}$ be the upper and lower $\frac{\alpha}{2}$ point by symmetry this indicates

$$\mathbb{P}_{\mu, \sigma^2}\left(-t_{\alpha/2, n-1} \leq \frac{\bar{X}-\mu}{S/\sqrt{n}} \leq t_{\alpha/2, n-1}\right) = 1-\alpha$$

$$\mathbb{P}_{\mu, \sigma^2}\left(\bar{X} - \frac{S}{\sqrt{n}} t_{\alpha/2, n-1} \leq \mu \leq \bar{X} + \frac{S}{\sqrt{n}} t_{\alpha/2, n-1}\right) = 1-\alpha$$

* **Def** : Confidence Set

A subset S is said to constitute a confidence set at confidence $(1-\alpha)$ if $\mathbb{P}(S(n) \in \Theta) \geq 1-\alpha \quad \forall \theta \in \Omega$.

* **Methods of finding Confidence Interval** :

Let θ be a parameter and T be a statistic based on a random sample of size n for the population. Most often it is possible to find a fuction $\Psi(T, \theta)$ whose distribution is independent of θ. Then, $\mathbb{P}(\Psi_{1-\alpha/2} < \Psi(T, \theta) < \Psi_{\alpha/2}) = 1-\alpha$ which can be reformulated as $\mathbb{P}(\theta_1(T) \leq \theta \leq \theta_2(T)) = 1-\alpha$ and the observed interval $[\theta_1(T), \theta_2(T)]$ is the CI with confidence level $(1-\alpha)$.

* **Example** :

- Find the CI of σ^2 :

Let $T(\sigma^2) = S^2$ and $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

$$\mathbb{P}\left(\chi^2_{1-\alpha/2, n-1} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2, n-1}\right) = 1-\alpha$$

$$\mathbb{P}\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}} \right) = 1-\alpha$$

Hence, the CI of σ^2 : $\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} , \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}} \right]$

- Find the CI of μ, σ^2 :

Boolsen Inequality : $\mathbb{P}(A \cap B) \geq \mathbb{P}(A) + \mathbb{P}(B) - 1$

Let $\alpha = \alpha_1 + \alpha_2$

$$\mathbb{P}\Big(\underbrace{\bar{X} - \frac{S}{\sqrt{n}} t_{\alpha/2,n-1} < \mu < \bar{X} + \frac{S}{\sqrt{n}} t_{\alpha/2,n-1}}_{A} \Big) + \mathbb{P}\Big(\underbrace{\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}}_{B} \Big) - 1$$

$= \mathbb{P}(A) + \mathbb{P}(B) - 1$

$= 1 - \alpha_1 + 1 - \alpha_2 - 1$

$= 1 - (\alpha_1 + \alpha_2)$

$= 1 - \alpha$

By Boolsen inequality : $\mathbb{P}(A \cap B) \geq 1 - \alpha$

CI of μ and σ^2 :

$$\mathbb{P}\left(\bar{X} - \frac{S}{\sqrt{n}} t_{\alpha/2,n-1} \leq \mu \leq \bar{X} + \frac{S}{\sqrt{n}} t_{\alpha/2,n-1} , \frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}} \right) = 1-\alpha$$

* **Asymptotic Confidence Interval** :

If n is large we can apply the CLT : $\sqrt{n}(\bar{X}-\mu) \longrightarrow N(0,\sigma^2)$ and S^2 is a consistent estimator of σ^2 $\frac{\sqrt{n}(\bar{X}-\mu)}{S} = \frac{\sigma}{S} \cdot \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \xrightarrow{n \to +\infty} N(0,1)$

$\mathbb{P}\left(\bar{X} - \frac{S}{\sqrt{n}} z_{\alpha/2} < \mu < \bar{X} + \frac{S}{\sqrt{n}} \right) = 1-\alpha$. As $n \to +\infty$ $t_{\alpha/2,n-1} \longrightarrow z_{\alpha/2}$.

* Example :

Let $X_1, \dots, X_n$ iid the asymptotic CI for λ.

$T(x) = \bar{X} \quad \Rightarrow \quad \mathbb{E}(\bar{X}) = \lambda \qquad Var(\bar{X}) = \frac{\lambda}{n}$

By CLT : $\frac{\sqrt{n}(\bar{X} - \lambda)}{\sqrt{\lambda}} \sim N(0,1)$ as $n \to +\infty$

Hence, $\mathbb{P}\left(-z_{\alpha/2} \leq \frac{\sqrt{n}(\bar{X} - \lambda)}{\sqrt{\hat{\lambda}}} \leq z_{\alpha/2}\right) = 1 - \alpha$

Therefore, the CI of λ at level $(1-\alpha)$: $\bar{X} \pm z_{\alpha/2}\sqrt{\frac{\lambda}{n}}$

* General Asymptotic CI for MLE :

We know that $\sqrt{n}(\hat{\theta} - \theta) \longrightarrow N(0, I(\theta)^{-1})$ as $n \to +\infty$. We estimate $I(\theta)$ by $I(\hat{\theta})$ by continuous mapping theorem $I(\hat{\theta}) \xrightarrow{\mathbb{P}} I(\theta)$, this interval is called Wald interval.

$$\sqrt{n\,I(\theta)}\,(\hat{\theta} - \theta) = \underbrace{\frac{\sqrt{I(\hat{\theta})}}{\sqrt{I(\theta)}}}_{\text{converges to 1 in } \mathbb{P}} \cdot \sqrt{n\,I(\theta)}(\hat{\theta} - \theta) \xrightarrow{n \to +\infty} N(0,1)$$

$$\Rightarrow \mathbb{P}\left(-z_{\alpha/2} \leq \sqrt{n\,I(\hat{\theta})}\,(\theta - \hat{\theta}) \leq z_{\alpha/2}\right) = 1 - \alpha$$

$$\Rightarrow \text{CI} : \quad \hat{\theta} \pm \frac{z_{\alpha/2}}{\sqrt{n\,I(\hat{\theta})}}$$

* Example :

Let $X_1, \dots, X_n \sim Ber(p)$ CI of $g(p) = \log\left(\frac{p}{1-p}\right)$.

- $\hat{p} = \bar{X} \Rightarrow g(\hat{p}) = \log\left(\frac{\bar{X}}{1-\bar{X}}\right)$

$\prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i}$

- $g'(p) = \frac{(1-p)}{(1-p)^2 \cdot p} = \frac{1}{p(1-p)}$

By the delta method : $\sqrt{n}\,(g(\hat{p}) - g(p)) \longrightarrow N\left(0, \frac{g'(p)^2}{I(p)}\right)$

$\sqrt{n}\,(g(\hat{p}) - g(p)) \longrightarrow N\left(0, \frac{1}{p(1-p)}\right)$

Hence, the CI of $g(\hat{p})$: $g(p) \pm z_{\alpha/2} \sqrt{\frac{1}{n\hat{p}(1-\hat{p})}}$ by the continuous mapping theorem and slutsky's lemma.

$\sqrt{np(1-p)}\,(g(\hat{p}) - g(p)) \longrightarrow N(0,1)$

$\sqrt{\frac{np(1-p)}{n\hat{p}(1-\hat{p})}}\,\sqrt{n\hat{p}(1-\hat{p})}\,(g(\hat{p}) - g(p)) \longrightarrow N(0,1)$

$\Rightarrow \sqrt{n\hat{p}(1-\hat{p})}\,(g(\hat{p}) - g(p)) \longrightarrow N(0,1)$

$\Rightarrow \mathbb{P}\left(-z_{\alpha/2} \le \sqrt{n\hat{p}(1-\hat{p})}\,(g(\hat{p}) - g(p)) \le z_{\alpha/2}\right) = 1 - \alpha$

* **Problem :**

Let $X_1, X_2, \dots, X_n \overset{iid}{\sim} Ber(p)$ be n tosses of a biased coin and let $\hat{p} = \bar{X}$. In this problem, we explore two different ways to construct a 95% CI for p.

By CLT : $\sqrt{n}\,(\hat{p} - p) \sim N(0, p(1-p))$ (1)

a) Use the plugin estimate $\hat{p}(1-\hat{p})$ to obtain a 95% CI for p.

$$\frac{\sqrt{n}\,(\hat{p}-p)}{\sqrt{\hat{p}(1-\hat{p})}} \longrightarrow N(0,1)$$

$$\mathbb{P}\left(-z_{0.025} \leq \frac{\sqrt{n}\,(\hat{p}-p)}{\sqrt{\hat{p}(1-\hat{p})}} \leq z_{0.025}\right) = 0.95$$

$$\mathbb{P}\left(\hat{p} - z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 0.95$$

$$p \in \left(\hat{p} \pm z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

b) Instead of using the plugin estimate $\hat{p}(1-\hat{p})$, note that (1) implies that for large n.

$$\mathbb{P}\left(-\sqrt{\hat{p}(1-\hat{p})}\, z_{\alpha/2} \leq \sqrt{n}\,(\hat{p}-p) \leq \sqrt{\hat{p}(1-\hat{p})}\, z_{\alpha/2}\right) \approx 1-\alpha$$

Solve this equation:

$\sqrt{n}\,(\hat{p}-p) = \sqrt{p(1-p)}\, z_{\alpha/2}$ for p in terms of $\hat{p}$ and

$\sqrt{n}\,(\hat{p}-p) = -\sqrt{p(1-p)}\, z_{\alpha/2}$ for p in terms of $\hat{p}$.

to obtain a different 95% CI for p.

$$n(\hat{p}-p)^2 \leq p(1-p)\, z^2_{\alpha/2}$$

$$\Rightarrow (n + z^2_{\alpha/2})\, p^2 - (2n\hat{p} + z^2_{\alpha/2})\, p + n\hat{p}^2 \leq 0 \quad , \text{ put } \alpha = 0.05$$

$$p \in \left(\frac{\hat{p} + \frac{z^2_{0.025}}{2n} \pm z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2_{0.025}}{4n^2}}}{1 + \frac{z^2_{0.025}}{n}}\right)$$

c) Perform a simulation study to determine the true coverages of the confidence intervals in parts (a) and (b) for the 9 combinations of sample size $n = 10, 40, 100$ and $p = 0.1, 0.3, 0.5$ (for each perform at least $B = 100,000$ simulations). In each simulation, you may estimate $\hat{p}$ directly from $\frac{1}{n} Bin(n, p)$. Report the simulated coverage levels in two tables and answer which CIs for p yield the true coverage close to 95% for small values of n?

* Bayesian Analysis:

So far we assumed θ is unknown but non-random quantity - it is actually some fixed value describing the true distribution of data and our goal was to estimate θ, "frequentist paradigm of S.I".

Now, in Bayesian Inference, we model θ as a random variable. The Bayesian Paradigm naturally incorporate our prior belief about the unknow parameter and update this belief based on observed data.

* Prior and Posterior Distribution:

Let X, Y be two RVs having PDF/PMF $f_{X,Y}(x, y)$.

Marginal: $$f_X(x) = \begin{cases} \int f_{X,Y}(x,y)\, dy, & \text{PDF} \\ \sum_{y \in S_Y} f_{X,Y}(x,y), & \text{PMF} \end{cases}$$

Conditional: $$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

$$f(x,y) = f(y|x) \cdot f(x) = f(x|y) f(y)$$

In Bayesian, before the data is observed, the unknown parameter is modelled as a random variable having a probability distribution, $f_\Theta(\theta)$ called prior distribution. This distribution represents our belief about the value of this parameter.

Instead of Y, we have the two RVs X and Θ :

$$f_{X,\Theta}(x,\theta) = f_{X|\Theta}(x,\theta)\, f_\Theta(\theta)$$

$$f_{\Theta|X}(\theta|x) = \frac{f_{X,\Theta}(x,\theta)}{f_X(x)}$$

$$\text{updated knowledge of } \Theta = \frac{f_{X|\Theta}(x,\theta)\, f_\Theta(x)}{\int f_{X,\Theta}(x,\theta)\, d\theta} = \frac{f_{X|\Theta}(x,\theta)\, f_\Theta(x)}{\int f_{X|\Theta}(x,\theta')\, f_\Theta(\theta')\, d\theta} \longrightarrow \text{Bayes Theorem}$$

(the denominator is the normalizing constant)

$$\propto f_{X|\Theta}(x,\theta)\, f_\Theta(x) \rightarrow \text{Prior Knowledge of } \Theta$$

(the observed data)

Posterior $\propto$ Likelihood . Prior

- Uniform $[0,1]$ (Uniform Prior) is similar to using a frequentist method.

* **Example** :

Let $P \in]0,1[$ be the probability of heads for a biased coin and let $X_1, \ldots, X_n$ be the outcome of n tosses of this coin. If we don't have any prior information about p, we might choose for its prior distribution of $Unif[0,1]$ having PDF $f_p(p) = 1_{[0,1]}$.

$$f_{x,p}(n_1,\dots,n_n,p) = f_{x|p}(n_1,\dots,n_n|p)\cdot f_p(p)$$

$$= \prod_{i=1}^{n} p^{n_i}(1-p)^{1-n_i}\,\mathbb{1}_{]0,1[}$$

$$= p^{\sum n_i}(1-p)^{n-\sum n_i}\,\mathbb{1}_{]0,1[}$$

$$f_x(n_1,\dots,n_n) = \int_0^1 p^s(1-p)^{n-s}\,dp \quad , \quad s = \sum_{i=1}^{n} n_i$$

$$\sim \text{Beta}(s+1, n-s+1)$$

$$f_{p|x}(p\,|\,n_1,\dots,n_n) = \frac{p^s(1-p)^{n-s}\mathbb{1}_{]0,1[}}{B(s+1,n-s+1)} \quad , \quad B(s+1,n-s+1) = \frac{\Gamma(s+1)\Gamma(n-s+1)}{\Gamma(n+2)}$$

$$p\,|\,n_1,\dots,n_n \sim \text{Beta}(s+1, n-s+1)$$

For logistic regression, Beta prior is used as it is between $[0,1]$

* Example :

Suppose now, we have a prior belief that p is close to $\frac{1}{2}$. For example $B(\alpha,\alpha)$ has the mean $\frac{\alpha}{\alpha+\alpha} = \frac{1}{2}$ and variance $\frac{1}{8\alpha+4}$. The constant α may be chosen depending on how confident we are, a priori, that p is $\frac{1}{2}$ ($\alpha = 1 \Rightarrow$ Unif $[0,1]$ prior, $\alpha > 1 \Rightarrow$ prior is more concentrated around $\frac{1}{2}$). Can you calculate the posterior distribution.

$$f_{x|p}(n_1,\dots,n_n\,|\,p) = \prod_{i=1}^{n} p^{n_i}(1-p)^{n_i} = p^s(1-p)^{n-s}$$

$$f_p(p) = \frac{1}{B(\alpha,\alpha)}\cdot p^{\alpha-1}\cdot(1-p)^{\alpha-1}$$

$$f_{p|x}(n_1,\dots,n_n\,|\,p) \propto p^{s+\alpha-1}\cdot(1-p)^{n+\alpha-s-1} \sim \text{Beta}(s+\alpha, n+\alpha-s)$$

If $\alpha = 1$, we go back to the other case.

* **Conjugate Prior** : When the posterior has the same distribution as the prior (Ex. 2 Beta) then this type of prior is called conjugate prior.

* **Example** :

Let $\lambda \in]0, +\infty[$ be the parameter of poisson model $X_1, \dots, X_n \overset{iid}{\sim} Pois(\lambda)$. Assume the prior for Λ is Gamma (α, β). Show that Gamma is a conjugate prior for Λ.

$$f_{X|\Lambda}(x, \lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda} \cdot \lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda} \cdot \lambda^{\sum x_i}}{\left(\prod_{i=1}^{n} x_i!\right)} \propto e^{-n\lambda} \cdot \lambda^{s}$$

$$f_{\Lambda}(\lambda) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \cdot \lambda^{\alpha-1} \cdot e^{-\beta\lambda} \propto \lambda^{\alpha-1} \cdot e^{-\beta\lambda}$$

$$f_{\Lambda|X}(\lambda | x) \propto \lambda^{s+\alpha-1} \cdot e^{-(\beta+n)\lambda} \sim \text{Gamma}(\alpha + s, \beta + n)$$

Hence, Λ is a conjugate prior.

* **Point and Interval Estimation** :

In many practical applications, we would like to have a single estimate $\hat{\theta}$ as well as interval describing uncertainty about θ.

Posterior Mean / Mode can be used as the estimate of $\hat{\theta}$. For the interval $\mathbb{P}(\theta \in I \mid X) = 1 - \alpha$ will give us Bayesian Credible interval.

$\hat{p} = \frac{n+\alpha}{n+2\alpha}$ is the point estimate for Ex. 2

We can use 0.05 and 0.95 quantile of Beta$(\alpha+s, n-s+\alpha)$ as 90% Bayesian credible interval for p.

Frequentist	Bayesian
Data are a repeatable random sample	Parameters are unknown and described probabilistically
Parameters are fixed	Analysis is done conditionally on the observed data i.e data is treated as fixed
Underlying parameters remain constant during this repeatable process	

* **Example** :

Sampling from an exponential distribution. Suppose that the distribution of the lifetime of fluorescent tubes of a certain type is the exponential distribution with parameter θ. Suppose that $x_1, \ldots, x_n$ is a random sample of lamps of this type. Assume that $\theta \sim \text{Gamma}(\alpha, \beta)$ for known α, β.

$$f_{X|\Theta}(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} \theta e^{-\theta x_i} = \theta^n \cdot e^{-\theta \sum x_i}, \quad s = \sum_{i=1}^{n} x_i$$

$$f_{\Theta}(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \theta^{\alpha-1} \cdot e^{-\beta\theta} \propto \theta^{\alpha-1} e^{-\beta\theta}$$

$$f_{\Theta|X}(\theta \mid x_1, \ldots, x_n) \propto \theta^{\alpha+n-1} e^{-(\beta+s)\theta}$$

$f_{\Theta|X}(\theta \mid n_1, \ldots, n_n) \propto \theta^{\quad} e^{\quad}$

Posterior $\Theta \mid X = n \sim \text{Gamma}(\alpha + n, \beta + s)$

* **Def** : Loss Function

A loss function is a real valued function of two variables $L(\theta, a)$ where $\theta \in \Omega$ and $a \in \mathbb{R}$.

Squared loss : $L(\theta, a) = (\theta - a)^2$

Absolute loss : $L(\theta, a) = |\theta - a|$

* **Def** : Bayes Estimator

$T(\underset{\sim}{X})$ is an estimator of θ and $f(.)$ is a prior PDF / PMF of $\theta \in \Omega$. Consider the problem of estimating θ without being able to observe the data. If the statistician chooses a particular estimate a, then the expected loss will be

$$\mathbb{E}(L(\theta, a)) = \int_{\Omega} L(\theta, a) \cdot f(\theta)\, d\theta$$

The statistician wishes to choose an estimate a for which the expected loss is minimum.

If the statistician observes the data $\underset{\sim}{x}$ then $f(. \mid n)$ denote the posterior PDF / PMF. The expected loss becomes :

$$\mathbb{E}(L(\theta, a \mid n)) = \int_{\Omega} L(\theta, a) \cdot f(\theta \mid n)\, d\theta$$

For each possible value x of X. Let $\delta(n)$ denote a value of the estimate

a for which the expected loss is minimum. Then, the function $\delta(\underset{\sim}{X})$ is called the Bayes estimator of θ.

* **Remark** : The Bayes estimator is an estimater chosen to minimize the posterior mean of some measure of how far the estimater is from the parameter.

* **Corollary** : Property of Posterior Mean (squared)

Let $\theta \in \Omega \subset \mathbb{R}$. Suppose the squared loss function is defined and we know the posterior mean i.e $\mathbb{E}(\theta | \underset{\sim}{X})$ is finite. Then, the Bayes estimater of θ is : $\delta(x) = \mathbb{E}(\theta | x)$

* **Example** :

The estimate of $\hat{p} = \frac{\alpha + s}{n + 2\alpha}$ for the second example by this corollary.

* **Corollary** : Property of Posterior Median (Absolute)

Let $\theta \in \Omega \subset \mathbb{R}$. Suppose the absolute loss function is defined and we know the posterior median is finite. Then, the Bayes estimater of θ is $\delta^*(x)$.

* **Example** :

Suppose $X_1, \dots, X_n$ be a random sample from $N(\theta, \sigma^2)$ where θ is unknown. Prior for $\theta \sim N(\mu_0, v_0^2)$. Find out the posterior distribution of $\theta \mid X_1 = x_1, \dots, X_n = x_n$?

$$f_{X|\theta}(x \mid \theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\theta)^2}{2\sigma^2}\right)$$

$$\propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i-\theta)^2\right)$$

$$\propto \exp\left(\frac{-1}{2\sigma^2}\left(\sum_{i=1}^{n} (x_i-\bar{x}) + n(\theta-\bar{x})^2\right)\right)$$

$$\propto \exp\left(\frac{-n}{2\sigma^2}(\theta-\bar{x})^2\right) \quad \text{because it is independent of } \theta$$

$$f_{\theta}(\theta) = \frac{1}{\sqrt{2\pi v_0^2}} \cdot \exp\left(\frac{-1}{2\sigma^2}(\theta-\mu_0)^2\right) \propto \exp\left(\frac{-1}{2\sigma^2}(\theta-\mu_0)^2\right)$$

$$f_{\theta|X}(\theta \mid x) \propto \exp\left(\frac{-n}{2\sigma^2}(\theta-\bar{x})^2 - \frac{1}{2v_0^2}(\theta-\mu_0)^2\right)$$

$$\propto \exp\left(\frac{-1}{2v_1^2}(\theta-\mu_1)^2\right)$$

Hence, $\theta \mid X_1 = x_1, \dots, X_n = x_n \sim N(\mu_1, v_1)$, μ_1 is the Bayes estimator

$$v_1^2 = \frac{\sigma^2 v_0^2}{\sigma^2 + n v_0^2} \qquad \mu_1 = \frac{\sigma^2 \mu_0 + n\theta_0^2 \bar{x}}{\sigma^2 + n v_0^2}$$

* Normal Approximation for large n :

Consider Bayesian inference applied with the prior $f_{\theta}(\theta)$ for a parametric model $f_{X|\theta}(x \mid \theta)$.

log likelihood : $\ell(\theta) = \sum_{i=1}^{n} \log\left(f_{X|\theta}(x \mid \theta)\right)$

Posterior : $f_{\theta|x}(\theta \mid x_1, \dots, x_n) = f_\theta(\theta)\, f_{x|\theta}(x_1, \dots, x_n \mid \theta)$

$= f_\theta(\theta) \cdot \exp(l(\theta))$

Applying Second order Taylor expansion of $l(\theta)$ around the MLE of $\theta = \hat{\theta}$:

$$l(\theta) \approx l(\hat{\theta}) + (\theta - \hat{\theta}) \cancel{l'(\hat{\theta})}^{\,0} + \frac{(\theta - \hat{\theta})^2}{2} l''(\hat{\theta})$$

$$\approx l(\hat{\theta}) - \frac{n}{2} (\theta - \hat{\theta})^2 I(\hat{\theta}) \qquad \text{for large } n$$

$$f_{\theta|x}(\theta \mid \underset{\sim}{x}) \propto \exp\left(\underbrace{l(\hat{\theta})}_{\text{function of data, independent of } \theta} - \frac{n}{2} (\theta - \hat{\theta})^2 I(\hat{\theta}) \right) f_\theta(\theta)$$

$$\propto \exp\left(-\frac{n}{2} (\theta - \hat{\theta})^2 I(\hat{\theta}) \right) f_\theta(\theta)$$

$$\propto \exp\left(\frac{-1}{2} \frac{(\theta - \hat{\theta})^2}{1/n I(\hat{\theta})} \right)$$

Hence, $f_{\theta|x}(\theta \mid \underset{\sim}{x}) \sim N\left(\hat{\theta}, \frac{1}{n I(\hat{\theta})}\right)$.

* **Remark** : To summarize the posterior mean for θ is, for large n, is approximately the MLE $\hat{\theta}$. Furthermore, a $100(1-\alpha)$ Bayesian Credible interval is approximately given by $\hat{\theta} \pm z_{\alpha/2} \frac{1}{\sqrt{n I(\hat{\theta})}}$ which is exactly the $100(1-\alpha)\%$ Wald CI for θ. In this sense, frequentist & Bayesian method yield similar inference for large n.